

WHAT IS CLAIMED IS:

- 1 1. A method of encoding input data within a system, wherein the input
2 data might include sequences of symbols that repeat in the input data or occur in other input
3 data encoded in the system, the method comprising:
 - 4 identifying, within a number of sequential input data symbols defined by an
5 offset and a window size, a fingerprint representation of the number of sequential input data
6 symbols;
 - 7 determining, from the fingerprint representation, whether the offset is to be
8 designated as a cut point;
 - 9 repeating the above steps of identifying and determining to arrive at a set of
10 cut points;
 - 11 segmenting the input data as indicated by the set of cut points;
 - 12 for each segment, determining whether the segment is to be a referenced
13 segment or an unreferenced segment;
 - 14 for each referenced segment, replacing the segment data of the referenced
15 segment with a reference label;
 - 16 for each referenced segment not already present in a persistent segment store,
17 storing a reference binding in the persistent segment store, wherein a reference binding
18 associates a referenced segment's data and its reference label;
 - 19 determining whether any sequence of segments is to be grouped as a reference
20 group;
 - 21 for each reference group, replacing the references in the group with a group
22 label; and
 - 23 for each reference group not already present in the persistent segment store,
24 storing a group reference binding in the persistent segment store, wherein a group reference
25 binding associates a reference group's references with its group label.
- 1 2. The method of claim 1, further comprising:
 - 2 recursively identifying groups of labels into higher level groups, wherein
3 groups of labels are one or more of groups of reference labels and groups of group labels;
 - 4 for each higher level group, replacing the higher level group with a group
5 label; and

6 for each higher level group not already present in the persistent segment store,
7 storing a group reference binding in the persistent segment store for the higher level group.

1 3. The method of claim 1, wherein the input data comprises payloads of
2 messages between clients and servers in a client-server network.

1 4. The method of claim 1, wherein the input data comprises portions of
2 files in an on-line backup system, further comprising representing files in the on-line backup
3 system as sequences of at least one of reference labels and group labels, and storing contents
4 of the persistent segment store as part of the on-line backup system.

1 5. The method of claim 1, wherein the input data comprises portions of
2 files in a file system, further comprising representing files in the file system as sequences of
3 at least one of reference labels and group labels and a segment store.

1 6. The method of claim 1, wherein the input data comprises portions of
2 files to be used in a file system, the method further comprising:

3 when storing a file to the file system, encoding it with at least one segment of
4 the file being represented as a segment referenced in the persistent segment store; and

5 when retrieving a file from the file system, caching the file in a local file store
6 as a decoded file, wherein each reference label and each group label is replaced with
7 corresponding segment data from the persistent segment store.